

Class 17: Analyzing Sequencing Data in the Cloud

Zoe Matsunaga (PID: A16853288)

Table of contents

Downstream Analysis	1
PCA	2
PC1 vs PC2 in base R	3
PCA Plots using ggplot2	3
PC1 vs PC2	3
PC1 vs PC3	4
PC2 vs PC3	5
Differential Expression Analysis	6

Downstream Analysis

```
library(tximport)

folders <- dir(pattern="SRR21568*")
samples <- sub("_quant", "", folders)
files <- file.path(folders, "abundance.h5" )
names(files) <- samples

txi.kallisto <- tximport(files, type = "kallisto", txOut = TRUE)
```

1 2 3 4

```
head(txi.kallisto$counts)
```

	SRR2156848	SRR2156849	SRR2156850	SRR2156851
ENST00000539570	0	0	0.00000	0
ENST00000576455	0	0	2.62037	0
ENST00000510508	0	0	0.00000	0
ENST00000474471	0	1	1.00000	0
ENST00000381700	0	0	0.00000	0
ENST00000445946	0	0	0.00000	0

```
colSums(txi.kallisto$counts)
```

SRR2156848	SRR2156849	SRR2156850	SRR2156851
2563611	2600800	2372309	2111474

```
sum(rowSums(txi.kallisto$counts)>0)
```

```
[1] 94561
```

```
to.keep <- rowSums(txi.kallisto$counts) > 0  
kset.nonzero <- txi.kallisto$counts[to.keep,]
```

```
keep2 <- apply(kset.nonzero,1,sd)>0  
x <- kset.nonzero[keep2,]
```

PCA

```
pca <- prcomp(t(x), scale=TRUE)
```

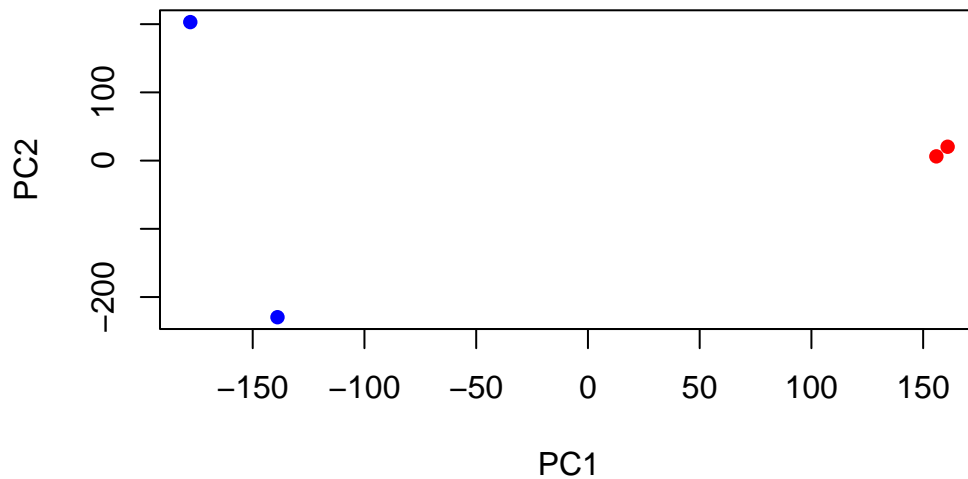
```
summary(pca)
```

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	183.6379	177.3605	171.3020	1e+00
Proportion of Variance	0.3568	0.3328	0.3104	1e-05
Cumulative Proportion	0.3568	0.6895	1.0000	1e+00

PC1 vs PC2 in base R

```
plot(pca$x[,1], pca$x[,2],  
     col=c("blue", "blue", "red", "red"),  
     xlab="PC1", ylab="PC2", pch=16)
```

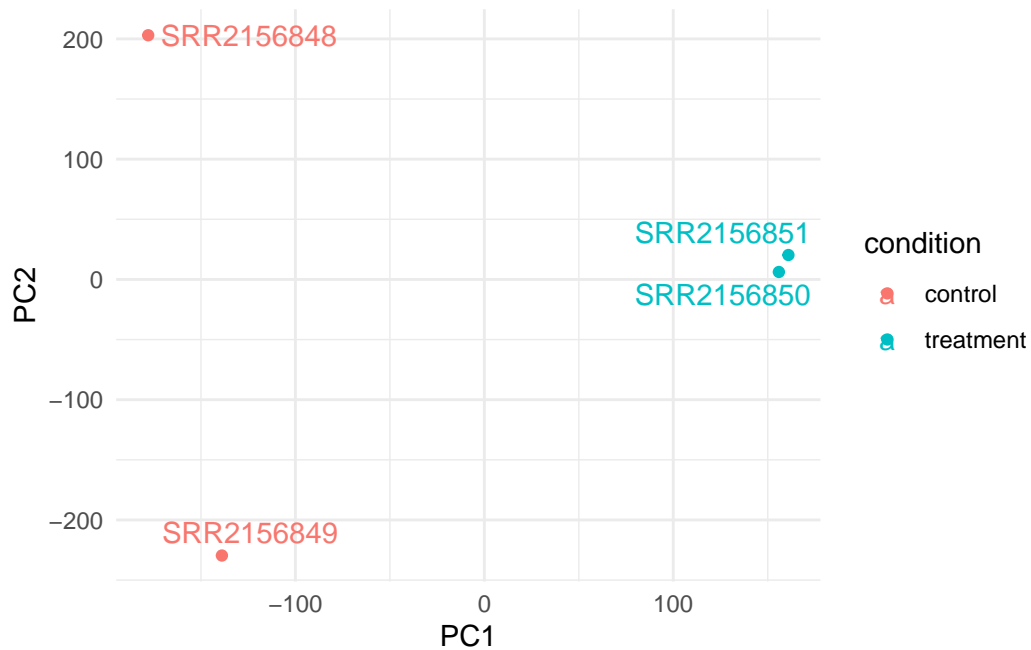


PCA Plots using ggplot2

```
pca.df <- data.frame(  
  sample = c("SRR2156848", "SRR2156849", "SRR2156850", "SRR2156851"),  
  PC1 = pca$x[,1],  
  PC2 = pca$x[,2],  
  PC3 = pca$x[,3],  
  condition = c("control", "control", "treatment", "treatment")  
)
```

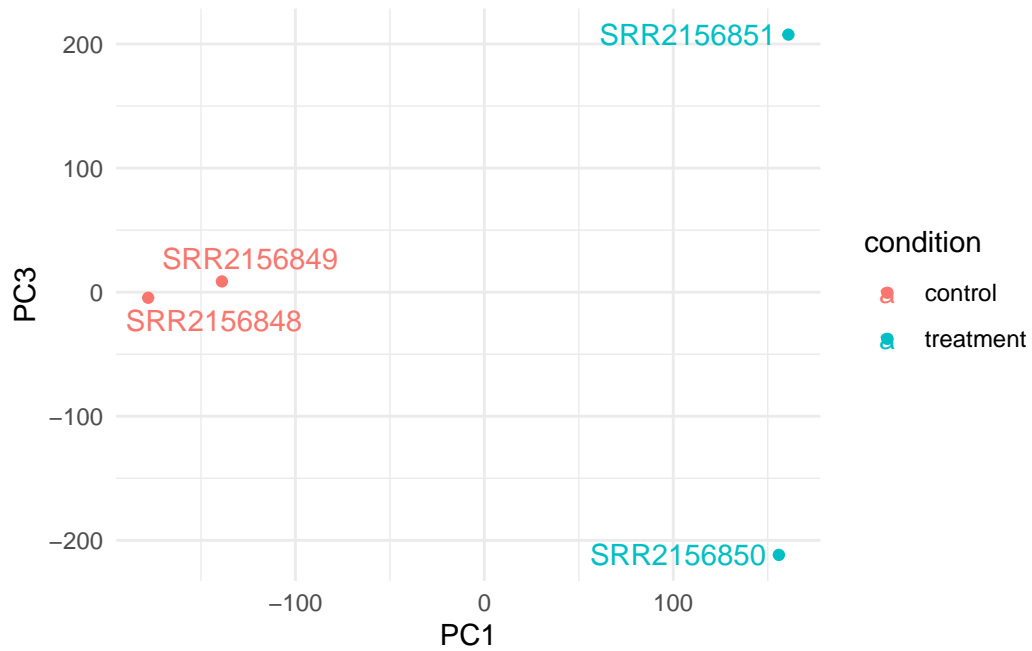
PC1 vs PC2

```
library(ggplot2)
ggplot(pca.df, aes(PC1, PC2, color = condition, label = sample)) +
  geom_point() +
  ggrepel::geom_text_repel() +
  labs(x = "PC1", y = "PC2") +
  theme_minimal()
```



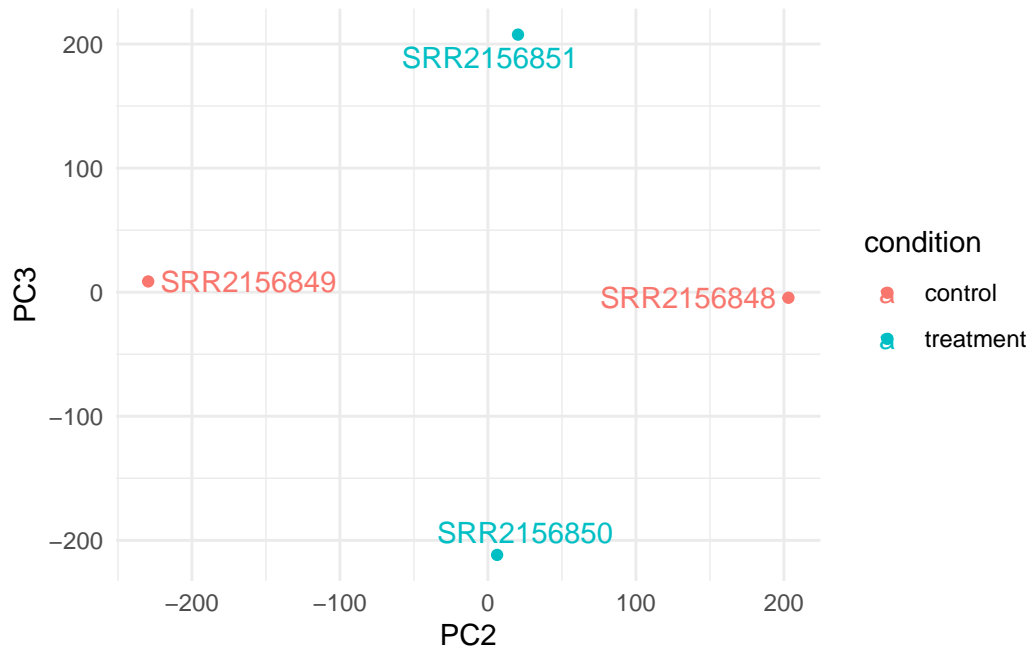
PC1 vs PC3

```
ggplot(pca.df, aes(PC1, PC3, color = condition, label = sample)) +
  geom_point() +
  ggrepel::geom_text_repel() +
  labs(x = "PC1", y = "PC3") +
  theme_minimal()
```



PC2 vs PC3

```
ggplot(pca.df, aes(PC2, PC3, color = condition, label = sample)) +  
  geom_point() +  
  ggrepel::geom_text_repel() +  
  labs(x = "PC2", y = "PC3") +  
  theme_minimal()
```



Differential Expression Analysis

```
library(DESeq2)
```

```
sampleTable <- data.frame(condition = factor(rep(c("control", "treatment"), each = 2)))
rownames(sampleTable) <- colnames(txi.kallisto$counts)
```

```
dds <- DESeqDataSetFromTximport(txi.kallisto,
                                sampleTable,
                                ~condition)
```

using counts and average transcript lengths from tximport

```
dds <- DESeq(dds)
```

estimating size factors

using 'avgTxLength' from assays(dds), correcting for library size

estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

-- note: fitType='parametric', but the dispersion trend was not well captured by the function: $y = a/x + b$, and a local regression fit was automatically substituted. specify fitType='local' or 'mean' to avoid this message next time.

final dispersion estimates

fitting model and testing

```
res <- results(dds)
head(res)
```

log2 fold change (MLE): condition treatment vs control

Wald test p-value: condition treatment vs control

DataFrame with 6 rows and 6 columns

	baseMean	log2FoldChange	lfcSE	stat	pvalue
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
ENST00000539570	0.000000	NA	NA	NA	NA
ENST00000576455	0.761453	3.155061	4.86052	0.6491203	0.516261
ENST00000510508	0.000000	NA	NA	NA	NA
ENST00000474471	0.484938	0.181923	4.24871	0.0428185	0.965846
ENST00000381700	0.000000	NA	NA	NA	NA
ENST00000445946	0.000000	NA	NA	NA	NA
	padj				
	<numeric>				
ENST00000539570	NA				
ENST00000576455	NA				
ENST00000510508	NA				
ENST00000474471	NA				
ENST00000381700	NA				
ENST00000445946	NA				